Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

GENOMIC CHARACTERIZATION OF SARS-CoV-2 AND ITS ASSOCIATION WITH CLINICAL OUTCOMES: A ONE-YEAR LONGITUDINAL STUDY OF THE PANDEMIC IN COLOMBIA

Ángela María Ruiz-Sternberg , Henry Mauricio Chaparro-Solano , Ludwig Luis Antonio Albornoz , Ángela María Pinzón-Rondón , Juan Mauricio Pardo-Oviedo , Nicolás Molano-González , Diego Andrés Otero-Rodríguez , Fabio Andrés Zapata-Gómez , Jubby Marcela Gálvez-Bermúdez

PII: S1201-9712(21)01220-0
DOI: https://doi.org/10.1016/j.ijid.2021.12.326
Reference: IJID 5883

To appear in: *International Journal of Infectious Diseases*

Received date: 23 August 2021
Revised date: 10 November 2021
Accepted date: 8 December 2021

Please cite this article as: Ángela María Ruiz-Sternberg , Henry Mauricio Chaparro-Solano , Ludwig Luis Antonio Albornoz , Ángela María Pinzón-Rondón , Juan Mauricio Pardo-Oviedo , Nicolás Molano-González , Diego Andrés Otero-Rodríguez , Fabio Andrés Zapata-Gómez , Jubby Marcela Gálvez-Bermúdez , GENOMIC CHARACTERIZATION OF SARS-CoV-2 AND ITS ASSOCIATION WITH CLINICAL OUTCOMES: A ONE-YEAR LONGITUDINAL STUDY OF THE PANDEMIC IN COLOMBIA, *International Journal of Infectious Diseases* (2021), doi: https://doi.org/10.1016/j.ijid.2021.12.326

HIGHLIGHTS

- Lineages B.1.621 and B.1.1.388 increased need for hospitalization and lethality
- Lineage B.1.621 was first detected in September 2020 worldwide
- B.1.621 became predominant in Colombia during the most serious outbreak
- Associations between clades and clinical outcomes diverged from past reported data

Essential title page information

Title: GENOMIC CHARACTERIZATION OF SARS-CoV-2 AND ITS ASSOCIATION WITH CLINICAL OUTCOMES: A ONE-YEAR LONGITUDINAL STUDY OF THE PANDEMIC IN COLOMBIA

Author names and affiliations:

- Ángela María Ruiz-Sternberg [a] (angela.ruiz@urosario.edu.co)
  - Carrera 24 # 63C-69, Bogotá, Colombia
- Henry Mauricio Chaparro-Solano [a,b,c] (mchaparro@labgencell.com)
  - Carrera 24 # 63C-69, Bogotá, Colombia
  - Calle 82 # 18 – 12, Bogotá, Colombia
  - Calle 24 # 29 – 45, Bogotá, Colombia
- Ludwig Luis Antonio Albornoz [d,e] (ludwig.albornoz@fvl.org.co)
  - Carrera 98 # 18 – 49, Cali, Colombia
  - Calle 18 # 122 – 135, Cali, Colombia
- Ángela María Pinzón-Rondón [a] (angela.pinzon@urosario.edu.co)
  - Carrera 24 # 63C-69, Bogotá, Colombia
- Juan Mauricio Pardo-Oviedo [a,c] (juan.pardo@urosario.edu.co)
  - Carrera 24 # 63C-69, Bogotá, Colombia
  - Calle 24 # 29 – 45, Bogotá, Colombia
- Nicolás Molano-González [a] (nicolas.molano@urosario.edu.co)
  - Carrera 24 # 63C-69, Bogotá, Colombia
- Diego Andrés Otero-Rodríguez [b] (dotero@labgencell.com)
  - Calle 82 # 18 – 12, Bogotá, Colombia
- Fabio Andrés Zapata-Gómez [b] (azapata@labgencell.com)
  - Calle 82 # 18 – 12, Bogotá, Colombia
- Jubby Marcela Gálvez-Bermúdez [b] (mgalvezb@labgencell.com)
  - Calle 82 # 18 – 12, Bogotá, Colombia

a: Clinical Investigation Group, Universidad del Rosario
b: Genuino Research Group, Gencell Pharma
c: Hospital Universitario Mayor – Méderi
d: Departamento de Patología y Medicina de Laboratorio, Fundación Valle del Lili
e: Facultad de Ciencias de la Salud, Universidad ICESI

Corresponding author:

- Ángela María Ruiz-Sternberg
- angela.ruiz@urosario.edu.co
- Carrera 24 # 63C - 69
- +57 1 2970200

ABSTRACT

Objectives: This study aims to explore the association between the molecular characterization of SARS-CoV-2 and disease severity on ambulatory and hospitalized patients in two main Colombian epicenters during the first year of the COVID-19 pandemic.

Methods:

We included 1000 patients with SARS-CoV-2 infection, collected clinical data from 997, and obtained 678 whole genome sequences by massively parallel sequencing. Bivariate, multivariate, and classification and regression tree analyses were run between clinical and genomic variables.

Results:

Age and infection with lineages B.1.1, B.1.1.388, B.1.523, and B.1.621 were related to lethality for patients 71-88 years old (OR: 6.048036; 95% CI 1.346567-32.92521, p-value: 0.01718674). The need for hospitalization was associated with higher age and comorbidities. For patients 38-51 years old infected with lineages A, B, B.1.1.388, B.1.1.434, B.1.153, B.1.36.10, B.1.411, B.1.471, B.1.558 or B.1.621, hospitalization rate increased significantly (OR 8.368427, 95% CI 2.573145-39.10672, p-value: 0.00012). Associations between clades and clinical outcomes diverged from previously reported data.

Conclusions:

Lineage B.1.621 increased the need for hospitalization and lethality. Our findings, plus the rapidly increasing prevalence in Colombia and other countries, suggest broadly considering it as a Variant of Interest. If associated disease severity is confirmed, possible designation as Variant of Concern could be entertained.

Keywords:

SARS-CoV-2; SARS-CoV-2 variants; COVID-19; Mortality; Hospitalization; High-Throughput Nucleotide Sequencing

INTRODUCTION

SARS-CoV-2, an RNA virus from the coronavirus family whose genome contains 29.8 Kb, has emerged as a new viral pathogen that causes COVID-19 respiratory disease. Due to its important transmission capabilities, this virus led to an unprecedented pandemic in human history, officially declared by the World Health Organization (WHO) in March 2020. By July 29, 2021, over 196 million cases have been reported and around 4 million deaths have been documented worldwide (Johns Hopkins Coronavirus Resource Center, 2021). Colombia (estimated 2020 population of 50.3 million) has been especially affected over time, with 4,877,323 cases and 123,781 deaths reported by the submission date of this article. Since the beginning of the pandemic, the country has faced three COVID-19 waves (July-August 2020; January 2021; and April 2021-present). The latter has been the most

aggressive, representing the second-highest worldwide number of daily new cases and deaths since May 2021 (*Coronavirus Colombia*, 2021) .

SARS-CoV-2 infection can have a wide spectrum of clinical outcomes, from asymptomatic infection to severe disease and death. Even though there are well-known sociodemographic and clinical risk factors related to COVID-19 clinical presentation, the influence of the viral mutational profile in infectivity and severity of the disease is yet to be fully elucidated (S.-W. Huang & Wang, 2021; Richardson et al., 2020). All viruses undergo genomic changes as they spread, but such variations mostly do not imply a structural or functional impact on protein translation (Peacock et al., 2021).

Since the complete sequence publication in December 2019, the SARS-CoV-2 genome has been thoroughly characterized, leading to the description of genes and regions that are important for host recognition and cellular entry, as well as immune response evasion. Different nomenclature systems based on the identification of mutation markers, like the Global Initiative on Sharing All Influenza Data -GISAID- that define 8 major clades (S, L, V, G, GH, GR, GV, and GRY), and/or genetic, epidemiological and geographical characteristics, such as the Phylogenetic Assignment Named Global Outbreak Lineages, -PANGO-, have been proposed (Elbe & Buckland-Merrett, 2017; Rambaut et al., 2020; Shu & McCauley, 2017). These systems are useful for tracking pandemic viral spread, allowing to explore a possible relation between novel genetic variants, lineages or clades, and disease severity,

even though by themselves these genetic variations may not suffice to explain viral phenotypic characteristics.

As a result of the PANGO system implementation and genomic surveillance programs established by countries around the world, an increasing number of lineages and variants have been described. Despite important efforts and investments made for the continuous sequencing of the SARS-CoV-2 genome, an insignificant proportion of variants have been recognized as epidemiologically or clinically relevant. These variants, called Variants of Concern (VOCs), Variants of Interest (VOIs), and Variants of High Consequence (VOHs), demand higher interest from governments and public health agencies since they contain changes that modify viral transmissibility, disease severity, and response to therapeutic and diagnostic tools (Janik et al., 2021). It has been hypothesized that these variants are the result of selective pressure due to changes in host immune characteristics as well as the development of new drugs, immunotherapy, and vaccines. The first VOC was the Alpha variant (B.1.1.7 lineage), identified in September 2020 in England (Galloway et al., 2021). Since then, other VOCs and VOIs have been reported, some displaying convergent mutations that confer the virus functional adaptive characteristics. These variants have become predominant and exhibit higher transmissibility and/or a significant impact on immunity and disease severity (*Tracking SARS-CoV-2 Variants*, 2021). Few studies have intended to longitudinally study the possible associations between the mentioned lineages, clades, or other classification systems, and clinical outcomes (Hamed et al., 2021; Lamptey et al., 2021; Nakamichi et al., 2021; Young et al., 2021).

This study aims to explore the association between the molecular characterization of SARS-CoV-2 and disease severity on ambulatory and hospitalized patients from two main cities in Colombia during the first year of the pandemic.

METHODS

Biological specimen collection, nucleic acid inactivation, and extraction

Informed consent was obtained from eligible patients. Prospective specimens: Confirmed SARS-CoV-2 respiratory tract specimens (nasopharyngeal aspirate or swab), RT-qPCR positive, were collected from patients recruited from two main pandemic epicenters in Colombia, at tertiary-care university hospitals and a molecular diagnostics laboratory. Retrospective specimens: RNA eluate or primary nasopharyngeal swabs/aspirates were obtained from biorepositories at the participating research centers. RT-qPCR negative samples were excluded. Demographic and clinical characteristics were collected in CASPIO (Caspio, Inc. Sunnyvale, California).

Viral RNA inactivation and extraction were performed on 0.2 ml aliquots of viral transport medium (primary sample swab specimens), or on 1 mL aliquots in sterile isotonic saline solution (primary aspirate samples). All specimens were heat-inactivated (56 ºC for 30 minutes) and manipulated under BSL level 2 conditions.

RNA extraction consisted of cell lysis, followed by bead binding to magnetic rods, RNA binding to beads, washing, and elution, to obtain 0.1 mL of the RNA eluate. Automated RNA extraction methods comprised ExiPrep™ 96 Viral DNA/RNA Kit, on ExiPrep™ 96 Lite

instrument (Bioneer Corp., Daejeon, Republic of Korea), MGIEasy Nucleic Acid Extraction Kit on MGISP-960 (MGI Tech Co. Ltd, Shenzhen, PRC), NucliSENS® Nucleic Acid Extraction Reagents on NucliSENS® easyMAG® (bioMérieux SA, Marcy l'Etoile, France), or MagNA Pure® Compact Nucleic Acid Isolation Kit I on MagNA Pure® Compact (Roche Diagnostics GmbH, Mannheim, Germany).

Statistical methods

Qualitative variables were reported as frequencies and percentages. Quantitative variables were reported as means and standard deviations or median and interquartile ranges depending on normality distribution.

To assess associations between viral genome characteristics (presence or absence of genetic variants, total number of variants per sample, total number of variants discriminated by gene and impact of the variant in the protein, PANGOLIN lineage, and GISAID clade) and death and need for hospitalization, we used the Kruskall-Wallis test or the Chi-Square of independence, correspondingly. In a second approach, we used the Classification and Regression Trees (Breiman et al., 2017) algorithm to find the relevant variables associated with death and hospitalization. This algorithm is useful since the number of covariates that can be included in the model is not limited, as is the case in more traditional approaches like logistic regression. We included as covariates the following: sex, age, number of comorbidities, asymptomatic status, BMI, and the aforementioned genetic characteristics.

The overall significance level was set at 5%. For all statistical analyses software R version 4.0.2 was used.

Phylogenetic analysis

All SARS-CoV-2 genomes were downloaded from SOPHiA™ DDM® bioinformatics software (SOPHiA Genetics Inc.). Fasta files were aligned to the reference genome, NC_045512, using MAFFT v7 software (Katoh et al., 2002). Next, a nucleotide substitution model was predicted using jMoldelTest v2.1.10 (Posada, 2008). Later, we constructed a maximum likelihood tree with IQ-TREE 2 software (Minh et al., 2020) using the GTR + Γ model and 1000 bootstrap replicates. Finally, each genome had a lineage assigned using the PANGOLIN webserver (Rambaut et al., 2020). Additionally, CoVsurver online server (*CoVsurver - CoronaVirus Surveillance Server*, 2021) was used for GISAID clade assignment.

The present project was approved by the IRB of Universidad del Rosario and of participating hospital Research Centers. All international and national bioethical principles and regulations for clinical investigation in human subjects are followed.

RESULTS

STUDY POPULATION CHARACTERISTICS

Demographic characteristics

Clinical and demographic information was obtained from 997 patients. The mean age was 50.6 years, 35% of participants were under 40 years and 33.7% were over 60 years. Sex distribution in the cohort was homogeneous. The ethnic majority accounted for 76.6% of the sample. Patients resided in Bogotá (62.2%) and in Cali (29%) (Table 1).

Clinical characteristics

At diagnosis, 90.7% of the patients had symptoms. Outpatients represented 50.8%, 33% were hospitalized, 9.2% received ICU support, and 6.9% died. The most frequent complications were respiratory (29.4%); symptoms were cough (54.5%), fatigue (52.4%), and fever (47.1%) (Table 2).

SARS-CoV-2 genetic characterization

We obtained 763 SARS-CoV-2 sequences, of which genomic coverage was >95% in most samples (63.4%). In 10.2% of cases, coverage was 75-95%; 25-75% and <25% coverage occurred in 13.4% and 13% of sequences, respectively.

A total of 2,715 single variants were identified: missense variants (54.1%), synonymous variants (37.75%) and Loss-of-Function -LoF- (3.6%); the remaining 4.7% included variants in the untranslated (3'UTR and 5'UTR) and intergenic regions, as well as in-frame, InDels, loss-of-start, and loss-of-stop codons variants.

Most genetic changes occurred in *ORF1ab* (63.5%), *S* (13.3%), and *N* (6.2%). When adjusted for kb, the highest rates were in *ORF8* (357.51 variants/kb), 3'UTR (218.34 variants/kb), and *N* (183.92 variants/kb); the lowest rates included *ORF1ab* (80.97 variants/kb), S (94.19 variants/kb), and M (101.64/kb).

Phylogenetic analysis

Due to poor genomic coverage, 85 sequences were discarded from phylogenetic analysis. Additionally, 5 samples did not pass the Chi-square test performed by Iqtree and were thus excluded. The maximum likelihood tree constructed yielded one major group (658 samples). The remaining 15 samples were clustered into 7 minor branches, more closely related to the original strain (Figure 1).

We identified 50 PANGO lineages, being B.1 most prevalent (45.0%) followed by B.1.111 (11.4%) and B.1.1.348 (9.7%). Interestingly, B.1.621, the so-called "Colombian Variant", was found in 7 cases (1.0%) (Table 3). Concerning GISAID clades, GH was predominant (48.1%); G, GR, and "other" clades were found in 24.8%, 19.8%, and 7.1% of cases, respectively. S and GRY clades were detected each in a single sample.

Demographic and clinical factors associated with mortality

A higher association with death was found in male patients over 60 years old and with several comorbidities. Multi-organic complications were associated with higher fatality. A distinct relationship was identified between mortality and the level of schooling: the death rate rose in patients with low educational levels (Table 4).

Hospitalization risk progressively increased as a function of age. Male sex, lower education level, and most comorbidities were associated with higher hospitalization requirements.

Interestingly, history of current or previous smoking was inversely related to hospitalization rate (Table 5).

Sociodemographic and clinical factors associated with the GISAID genetic clades

G and GR clades predominated in residents of Bogotá and GH in residents of Cali. Associations between clades and comorbidities were identified: GH and “Others” with diabetes mellitus type 2 (DM2), malignancy, and obesity, while clade G with nephropathy. Symptoms such as cough and fatigue were seen more frequently in patients infected with G clade, while nasal congestion with GR0 and respiratory and renal complications with clades grouped as “Other”.

Need for hospitalization and ICU care were associated with clade G and “Other”, while clades GH and GR predominated in outpatients. Sequences classified as “Others” were more frequent in patients over 59 years old (Table 6).

GISAID clades composition changed throughout the study time window (Figure 2). Clade G and “others” were more abundant during the first months. The prevalence of clade GR grew from 4% initially to 26% at the study conclusion. Similarly, clade GH prevalence rose from 21% to 64%.

Regarding the results of the decision tree, Figure 3 depicts the most important variables determining patient mortality. Of all variables considered, only age (88 years and older),

and infection by viral lineages B.1.1, B.1.1.388, B.1.523, B.1.621 in patients 71-88 years old, associated with lethality (OR: 6.048036; 95% CI 1.346567 - 32.92521, p-value: 0.01718674).

Similarly, for hospitalization, the tree showed that patients 59 years and older, and patients between 59 and 51 years with comorbidities, presented higher hospitalization rates. It is noteworthy that in patients 38 to 51 years old, infection by certain viral lineages associated with higher hospitalization rates, namely A, B, B.1.1.388, B.1.1.434, B.1.153, B.1.36.10, B.1.411, B.1.471, B.1.558 or B.1.621, (OR 8.368427, 95% CI 2.573145 - 39.10672, p-value: 0.00012) (Figure 4).

Discussion

Despite the great number of genetic variants and lineages identified among the obtained sequences, and consistent with previously published evidence (Peacock et al., 2021), just two were associated with both increased hospitalization and fatality rates. Firstly, B.1.621, recently labeled as a "Colombian lineage", was detected in Colombia on January 11, 2021, by the National Institute of Health (INS) (Laiton-Donato et al., 2021). Interestingly, in the present study we report the detection of this lineage in a sample from September 2020, collected in Bogotá. B.1.621 has disseminated nationally with significant acceleration since March 2021, reaching 26% accumulated prevalence, and peaking with 71% prevalence (7-day rolling average) by the end of July 2021, according to data uploaded to GISAID by the Colombian National Genomic Surveillance Program, as per the online data aggregating

and lineage/mutation tracker outbreak.info (Elbe & Buckland-Merrett, 2017; Shu & McCauley, 2017). B.1.621 was estimated to be 52.7% of circulating lineages between April and June 2021 (http://www.ins.gov.co/BibliotecaDigital/Estrategia-de-caracterizacion-genomica-SARS-CoV2_Colombia.pdf). The B.1.621 lineage has been detected in 28 countries, prompting active monitoring by the WHO since May 26, 2021 (*Tracking SARS-CoV-2 Variants*, 2021); Public Health England upgraded it from active monitoring to Variant Under Investigation (VUI-21JUL-01) on July 21, 2021 (*SARS-CoV-2 Variants of Concern and Variants under Investigation*, 2021); the European Centre for Disease Prevention and Control (ECDC) classified it on July 29, 2021, as a VOI (*SARS-CoV-2 Variants of Concern as of 5 August 2021*, 2021). By the date of submission of this paper, the U.S. CDC had not yet designated it, although it had been detected in 28 US states (Elbe & Buckland-Merrett, 2017; Shu & McCauley, 2017).

We adhere to the growing concern over B.1.621 as reflected by its designation as a VOI by the ECDC. Admittedly, and previous to our findings, there was insufficient real-world, experimental, or model-based evidence as pertains to the impact of B.1.621. Such lack of evidential substantiation probably reflects that B.1.621 worldwide prevalence is very low, reportedly less than 0.5% (Elbe & Buckland-Merrett, 2017; Shu & McCauley, 2017). There may be an under-representation of B.1.621 as a result of insufficient detection and report by genomic surveillance, although Colombian sequencing efforts have provided over 23% of the global GISAID variant sequences for B.1.621. Lineage B.1.621 seems to have already displayed heightened ease of transmission, by playing a central role in variant prevalence,

considering the current and most serious national COVID-19 outbreak wave in Colombia. The latter is coupled with case restriction to well-delimited regions in a specific geography, allowing to fulfill criteria for VOI consideration (Janik et al., 2021).

As pertains to genomic characterization, B.1.621 lineage shows the accumulation of substitutions including I95I, Y144T, Y145S in the N terminal domain; R346K, E484K, N501Y in the receptor-binding domain; P681H in the S1/S2 cleavage site; insertion 146N in the spike protein (Laiton-Donato et al., 2021), and K417N S gene mutation (*SARS-CoV-2 Variants of Concern and Variants under Investigation*, 2021).

A potential designation as VOC requires, in addition to the attributes shared with VOIs, an evident rise in disease severity. In fact, through the decision tree analysis, our findings shed light on the impact of lineage B.1.621, give support to its consideration as a possible VOC, and could prompt action so that it may become a target of strengthened public health measures and focused genomic surveillance. Although in Colombia the Delta Variant (B.1.617.2 lineage) was first identified by the end of July 2021, up to date it does not have the dominant prevalence it displays in many parts of the world.

We report that lineage B.1.1.388 is associated with a higher hospitalization rate and lethality. B.1.1.388 was initially and almost exclusively reported in Colombia (until recently in Ecuador and Spain), triggering PANGO to label it as another “Colombian

lineage" (Rambaut et al., 2020), and displays several distinctive substitutions that have not been designated as neither VOI nor VOC by the date of this paper's submission.

A high percentage of patients had symptoms (90.7%), mostly with influenza-like illness. Clinical presentation severity allowed for outpatient management in 50.8% of cases; 33% required hospitalization and 9.2% ICU admission; lethality was 6.9%. Admission rates to ICU coincide with reports in the literature (5%-32%) (Guan et al., 2020; C. Huang et al., 2020).

In agreement with others, pulmonary complications presented predominantly (29.4%). Other systems had minor involvement, mainly associated with the multisystemic impact of the disease. As in most series, conditions most frequently associated with hospitalization or ICU admission were age over 60 years, male sex, hypertension, cardiovascular disease, nephropathy, obesity, or thyroid disease (Chen et al., 2020; Guan et al., 2020; C. Huang et al., 2020; Richardson et al., 2020; Wu et al., 2020).

Befitting the literature, age was the most important variable associated with hospitalization requirement (21% <38 years vs 77% >59 years). Remarkably, lineages A, B, B.1.1.388, B.1.1.434, B.1.153, B.1.36.10, B.1.411, B.1.471, B.1.558, or B.1.621, increased the likelihood of hospitalization (82% vs. 35%) in relatively young patients.

For mortality, the behavior was similar (3% <71 years vs 26% >88 years old). The presence of lineages B.1.1, B.1.1.388, B.1.523, or B.1.621 was associated with increased mortality (62% vs. 21%).

In the bivariate analysis, we found a significant association between mortality and clades G and “others” possibly because patients over 60 years old and with comorbidities were overrepresented in these two clades, leading to possible bias. Furthermore, in the first phase of the study, a higher proportion of G and GH clades were identified, as opposed to GH and “others” in the second. This association between mortality and clades G and “others” was no longer seen on the multivariate analysis.

Several studies have explored associations between clinical outcomes and SARS-CoV-2 clades, and findings have been broadly divergent. Hamed et al, found that GH and GR clades were associated with severe/deceased outcomes, while S, G, and GV were associated with mild/asymptomatic cases. Clades L and V showed no significant statistical association (Hamed et al., 2021).

Young et al, showed clade L/V to have a significant association with severity and a more intense systemic inflammatory response, while clade G was not associated with higher severity or transmissibility (Young et al., 2021).

Nakamichi et al, explored the association between genetic variants and hospitalization and mortality due to SARS-CoV-2 infection, designating two clear clades from hierarchical clustering of the sequence variants. Clade 2, predominantly composed of S clade, showed a trend toward poorer clinical outcomes compared with Clade 1, predominantly constituted by the GH clade (Nakamichi et al., 2021).

Taxonomic classification into clades provides for a relatively coarse characterization, possibly lacking sufficient granularity to do clinical correlation because clades are constituted by lineages designated or not as VOCs or VOIs. Additionally, clade composition could be modified over time, depending on the identification of new lineages and the understanding of the clinical impact on COVID-19 severity that previously designated lineages could have. In this sense, the allocation of a VOC or VOI in a specific clade could give a false perception that the clade by itself is the variable associated with higher severity or transmissibility, in place of lineage, which may indeed be what relates with a worse clinical severity outcome. For instance, the Alpha variant belongs to the GRY clade (previously called GR); Beta, Epsilon and Iota to GH; Gamma, Zeta, Theta and Lambda to GR; and Delta, Eta, and Kappa to G. Such a wide distribution of VOIs and VOCs hinders the exploration of possible associations (*Tracking SARS-CoV-2 Variants*, 2021). Finally, these are dynamic lineage groupings prone to reclassification and reallocation.

Conclusions

In this study, we have described the association between SARS-CoV-2 lineages and the rates of patient hospitalization and fatality. Our findings, in context with that of others, make plausible the consideration of lineage B.1.621 as a VOI. In our view, VOI designation of lineage B.1.621 merits consideration due to the fixation and significant increase in the detection frequency over a relatively short interval, and because of the high detection rate within the protracted third wave of SARS-CoV-2 infection in Colombia, which was the third-largest COVID-19 caseload in Latin America, twelfth 12 globally. As the disease severity for this lineage is better characterized in further studies, a possible designation as a VOC could be entertained.

This is a cohort study, viewed in contrast to GISAID data and ecological studies. We suggest that in public databases such as GISAID, clinical information associated with the sequence data would be beneficial if made available, to foment a timelier association of genomic data with clinical variables. This may prompt a more expedient consideration for variant classification as VOCs or VOIs, in turn, triggering strict surveillance in terms of public health and other policies related to the management of the pandemic.

Our study included patients with follow-up until clinical outcome definitions were met, ensuring the fidelity of information collected. In addition, the twelve-month temporal coverage sheds light on the evolution of SARS-CoV-2 and the dynamics related to the introduction of the pathogen from other countries. Preanalytical specimen management

included a unique platform for obtaining sequences and automated library preparation, thus controlling for cross-contamination and operator-dependent error.

This study has limitations. First, it uses a convenience sample, which limits its generalizability. While participants come from the largest and third-largest cities in Colombia, admittedly the major national outbreak epicenters especially in the first half of 2020, Caribbean coastal populations were excluded, in whom most of the emerging B.1.621 variant cases are detected. Second, we had a 67% rate of successful sequencing: failure was mostly due to lower viral sample contents and/or RNA degradation. Third, the study design was ambispective: recall bias may have affected the accuracy of symptoms information provided in retrospective cases. Nonetheless, good agreement with the literature leads us to infer that recall bias is probably minor. Finally, patient recruitment concluded shortly before the third, and as of yet most serious, pandemic wave in Colombia, where B.1.621 lineage detection soared, thus explaining its relatively low frequency in our cohort study.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We thank Lina Marcela Méndez Castillo, Andrés Felipe Torres Gómez and Danyela Faisury Valero Rubio for supporting the laboratory procedures, and Andrea del Pilar Hernández Rodríguez, Andrés Felipe Patiño Aldana, Liseth Belinda Cifuentes Saenz, Natalia Andrea Pedraza López, Mateo Andrés Díaz Quiroz, Kevin Daniel Rivera Mendoza, Isabella Caicedo, María Alejandra Urbano, Alejandra Hidalgo, and all other members of the Clinical Investigation Group that supported the collection of clinical data and development of this research.

Funding: This work was supported by governmental funding from the Colombian Ministry of Science, Technology and Innovation -Ministerio de Ciencia, Tecnología e Innovación- (Grant number 366-2020).

Ethical approval

The present project was approved by the IRB of Universidad del Rosario and of participating hospital Research Centers. All international and national bioethical principles and regulations for clinical investigation in human subjects are followed.

References

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (2017). Classification And Regression Trees. Routledge. https://doi.org/10.1201/9781315139470

Chen, N., Zhou, M., Dong, X., Qu, J., Gong, F., Han, Y., Qiu, Y., Wang, J., Liu, Y., Wei, Y., Xia, J., Yu, T., Zhang, X., & Zhang, L. (2020). Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: A descriptive study. Lancet (London, England), 395(10223), 507–513. https://doi.org/10.1016/S0140-6736(20)30211-7

Coronavirus Colombia. (2021). Recuperado el 11 de junio de 2021, de

https://www.ins.gov.co/Noticias/paginas/coronavirus.aspx

COVID-19 Map. (2021). Johns Hopkins Coronavirus Resource Center. Recuperado el 11 de junio de 2021, de https://coronavirus.jhu.edu/map.html

CoVsurver—CoronaVirus Surveillance Server. (2021). Recuperado el 11 de junio de 2021, de https://mendel3.bii.a-star.edu.sg/METHODS/corona/beta/indexAnno2.html

Elbe, S., & Buckland-Merrett, G. (2017). Data, disease and diplomacy: GISAID's innovative contribution to global health. Global Challenges (Hoboken, NJ), 1(1), 33–46. https://doi.org/10.1002/gch2.1018

Galloway, S. E., Paul, P., MacCannell, D. R., Johansson, M. A., Brooks, J. T., MacNeil, A., Slayton, R. B., Tong, S., Silk, B. J., Armstrong, G. L., Biggerstaff, M., & Dugan, V. G. (2021). Emergence of SARS-CoV-2 B.1.1.7 Lineage—United States, December 29, 2020-January 12, 2021. MMWR. Morbidity and Mortality Weekly Report, 70(3), 95–99. https://doi.org/10.15585/mmwr.mm7003e2

Guan, W.-J., Ni, Z.-Y., Hu, Y., Liang, W.-H., Ou, C.-Q., He, J.-X., Liu, L., Shan, H., Lei, C.-L., Hui, D. S. C., Du, B., Li, L.-J., Zeng, G., Yuen, K.-Y., Chen, R.-C., Tang, C.-L., Wang, T., Chen, P.-Y., Xiang, J., … China Medical Treatment Expert Group for Covid-19. (2020). Clinical Characteristics of Coronavirus Disease 2019 in China. The New England Journal of Medicine, 382(18), 1708–1720. https://doi.org/10.1056/NEJMoa2002032

Hamed, S. M., Elkhatib, W. F., Khairalla, A. S., & Noreddin, A. M. (2021). Global dynamics of SARS-CoV-2 clades and their relation to COVID-19 epidemiology. Scientific Reports, 11(1), 8435. https://doi.org/10.1038/s41598-021-87713-x

Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., Zhang, L., Fan, G., Xu, J., Gu, X., Cheng, Z., Yu, T., Xia, J., Wei, Y., Wu, W., Xie, X., Yin, W., Li, H., Liu, M., … Cao, B. (2020). Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet (London, England), 395(10223), 497–506. https://doi.org/10.1016/S0140-6736(20)30183-5

Huang, S.-W., & Wang, S.-F. (2021). SARS-CoV-2 Entry Related Viral and Host Genetic Variations: Implications on COVID-19 Severity, Immune Escape, and Infectivity. International Journal of Molecular Sciences, 22(6). https://doi.org/10.3390/ijms22063060

Janik, E., Niemcewicz, M., Podogrocki, M., Majsterek, I., & Bijak, M. (2021). The Emerging Concern and Interest SARS-CoV-2 Variants. Pathogens (Basel, Switzerland), 10(6), 633. https://doi.org/10.3390/pathogens10060633

Katoh, K., Misawa, K., Kuma, K., & Miyata, T. (2002). MAFFT: A novel method for rapid

multiple sequence alignment based on fast Fourier transform. Nucleic Acids Research, 30(14), 3059–3066. https://doi.org/10.1093/nar/gkf436

Laiton-Donato, K., Franco-Muñoz, C., Álvarez-Díaz, D. A., Ruiz-Moreno, H. A., Usme-Ciro, J. A., Prada, D. A., Reales-González, J., Corchuelo, S., Herrera-Sepúlveda, M. T., Naizaque, J., Santamaría, G., Rivera, J., Rojas, P., Ortiz, J. H., Cardona, A., Malo, D., Prieto-Alvarado, F., Gómez, F. R., Wiesner, M., … Mercado-Reyes, M. (2021). Characterization of the emerging B.1.621 variant of interest of SARS-CoV-2. Infection, Genetics and Evolution, 105038. https://doi.org/10.1016/j.meegid.2021.105038

Lamptey, J., Oyelami, F. O., Owusu, M., Nkrumah, B., Idowu, P. O., Adu-Gyamfi, E. A., Czika, A., El-Duah, P., Yeboah, R., Sylverken, A., Olasunkanmi, O. I., Owusu-Dabo, E., Drosten, C., & Adu-Sarkodie, Y. (2021). Genomic and epidemiological characteristics of SARS-CoV-2 in Africa. PLoS Neglected Tropical Diseases, 15(4), e0009335. https://doi.org/10.1371/journal.pntd.0009335

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A., & Lanfear, R. (2020). IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Molecular Biology and Evolution, 37(5), 1530–1534. https://doi.org/10.1093/molbev/msaa015

Nakamichi, K., Shen, J. Z., Lee, C. S., Lee, A., Roberts, E. A., Simonson, P. D., Roychoudhury, P., Andriesen, J., Randhawa, A. K., Mathias, P. C., Greninger, A. L., Jerome, K. R., & Van Gelder, R. N. (2021). Hospitalization and mortality associated with SARS-CoV-2 viral clades in COVID-19. Scientific Reports, 11(1), 4802. https://doi.org/10.1038/s41598-021-82850-9

Peacock, T. P., Penrice-Randal, R., Hiscox, J. A., & Barclay, W. S. (2021). SARS-CoV-2 one year on: Evidence for ongoing viral adaptation. The Journal of General Virology, 102(4). https://doi.org/10.1099/jgv.0.001584

Posada, D. (2008). jModelTest: Phylogenetic model averaging. Molecular Biology and Evolution, 25(7), 1253–1256. https://doi.org/10.1093/molbev/msn083

Rambaut, A., Holmes, E. C., O'Toole, Á., Hill, V., McCrone, J. T., Ruis, C., du Plessis, L., & Pybus, O. G. (2020). A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nature Microbiology, 5(11), 1403–1407. https://doi.org/10.1038/s41564-020-0770-5

Richardson, S., Hirsch, J. S., Narasimhan, M., Crawford, J. M., McGinn, T., Davidson, K. W., the Northwell COVID-19 Research Consortium, Barnaby, D. P., Becker, L. B., Chelico, J. D., Cohen, S. L., Cookingham, J., Coppa, K., Diefenbach, M. A., Dominello, A. J., Duer-Hefele, J., Falzon, L., Gitlin, J., Hajizadeh, N., … Zanos, T. P. (2020). Presenting Characteristics, Comorbidities, and Outcomes Among 5700 Patients Hospitalized With COVID-19 in the New York City Area. JAMA, 323(20), 2052–2059. https://doi.org/10.1001/jama.2020.6775

SARS-CoV-2 variants of concern and variants under investigation. (2021). 55.

SARS-CoV-2 variants of concern as of 5 August 2021. (2021). European Centre for Disease Prevention and Control. Recuperado el 17 de agosto de 2021, de https://www.ecdc.europa.eu/en/covid-19/variants-concern

Shu, Y., & McCauley, J. (2017). GISAID: Global initiative on sharing all influenza data - from vision to reality. Euro Surveillance: Bulletin Europeen Sur Les Maladies Transmissibles = European Communicable Disease Bulletin, 22(13). https://doi.org/10.2807/1560-7917.ES.2017.22.13.30494

Tracking SARS-CoV-2 variants. (2021). Recuperado el 17 de agosto de 2021, de https://www.who.int/emergencies/emergency-health-kits/trauma-emergency-surgery-kit-who-tesk-2019/tracking-SARS-CoV-2-variants

Wu, C., Chen, X., Cai, Y., Xia, J., Zhou, X., Xu, S., Huang, H., Zhang, L., Zhou, X., Du, C., Zhang, Y., Song, J., Wang, S., Chao, Y., Yang, Z., Xu, J., Zhou, X., Chen, D., Xiong, W., … Song, Y. (2020). Risk Factors Associated With Acute Respiratory Distress Syndrome and Death in Patients With Coronavirus Disease 2019 Pneumonia in Wuhan, China. JAMA Internal Medicine, 180(7), 934–943. https://doi.org/10.1001/jamainternmed.2020.0994

Young, B. E., Wei, W. E., Fong, S.-W., Mak, T.-M., Anderson, D. E., Chan, Y.-H., Pung, R., Heng, C. S., Ang, L. W., Zheng, A. K. E., Lee, B., Kalimuddin, S., Pada, S., Tambyah, P. A., Parthasarathy, P., Tan, S. Y., Sun, L., Smith, G. J., Lin, R. T. P., … Lee, V. J. (2021). Association of SARS-CoV-2 clades with clinical, inflammatory and virologic outcomes: An observational study. EBioMedicine, 66, 103319. https://doi.org/10.1016/j.ebiom.2021.103319

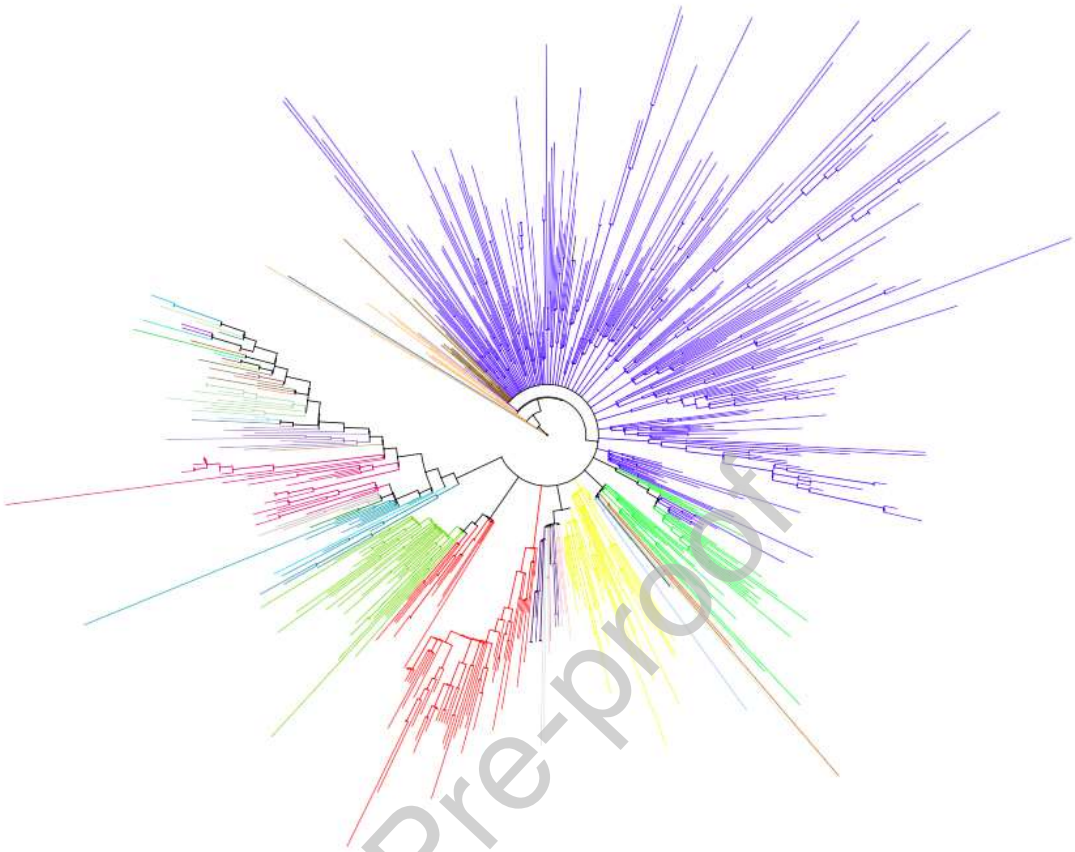

Figure 1. Phylogenetic tree of 673 Colombian samples. Most sequences clearly cluster together while the remaining 15 lies in different branches more closely related to the Wuhan strain. Most frequent lineages are labeled with the following colors: B.1 blue, B.1.111 red, B.1.1.348 yellow, B.1.1 green, B.1.153 olive green, B.1.420 pink. Other colors are described in HTML code in Table 1 supplementary material.

(This figure should be colored).

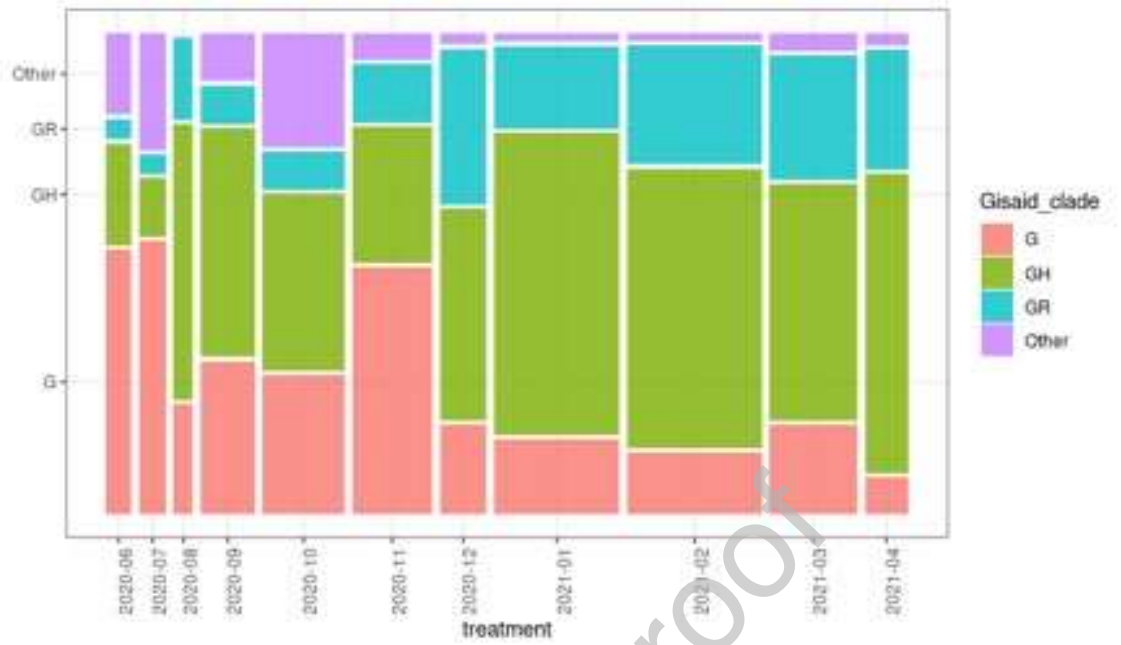

Figure 2. Mosaic plot showing the monthly distribution of GISAID clades during the study time window.

(This figure should be colored).

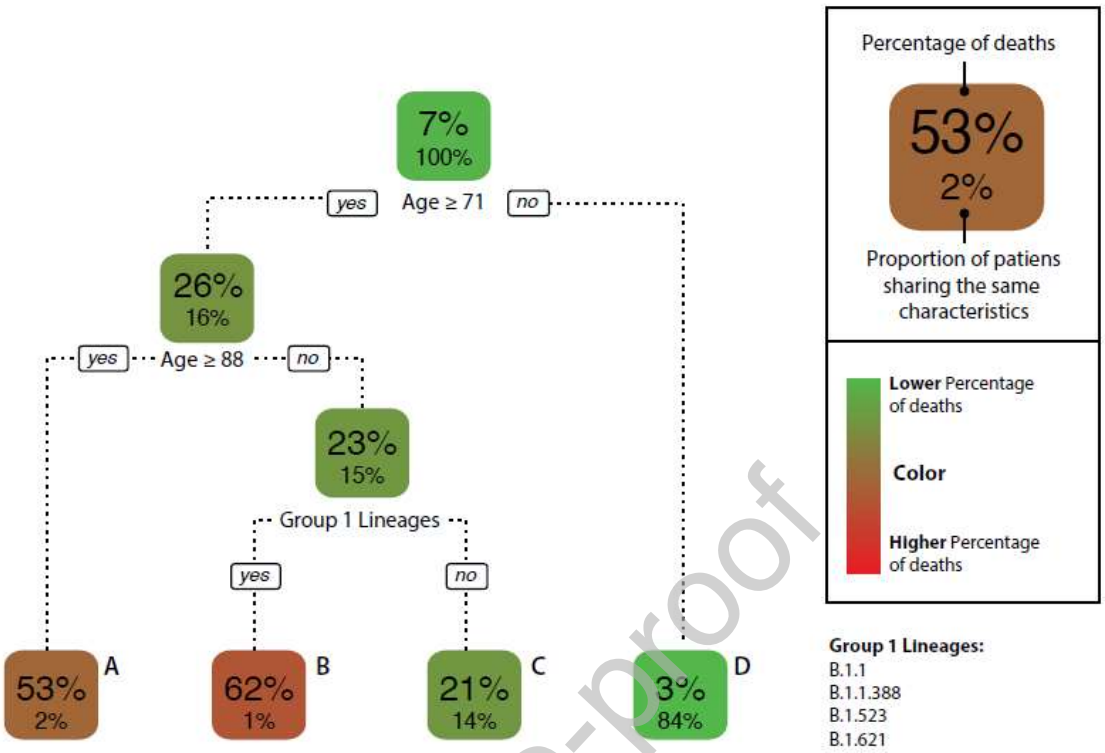

Figure 3. Classification and regression tree for mortality

Two variables (age, and four lineages) were associated with higher mortality. This tree identifies patients with commonalities, classified into four subgroups: A) Patients over 88 years old, with mortality of 53%, (2% of the total study sample). B) Patients 71 to 88 years old, who presented with either of the following four lineages: B.1.1, B.1.1.388, B.1.523, or B.1.621. In this group, 62% of patients were deceased (1% of the sample). C) Patients between 71 and 88 years old, presented with lineages different from the four described above, and had 21% mortality (14% of the sample). D) Patients under 71 years old, who had 3% mortality (84% of the total sample).

(This figure should be colored).

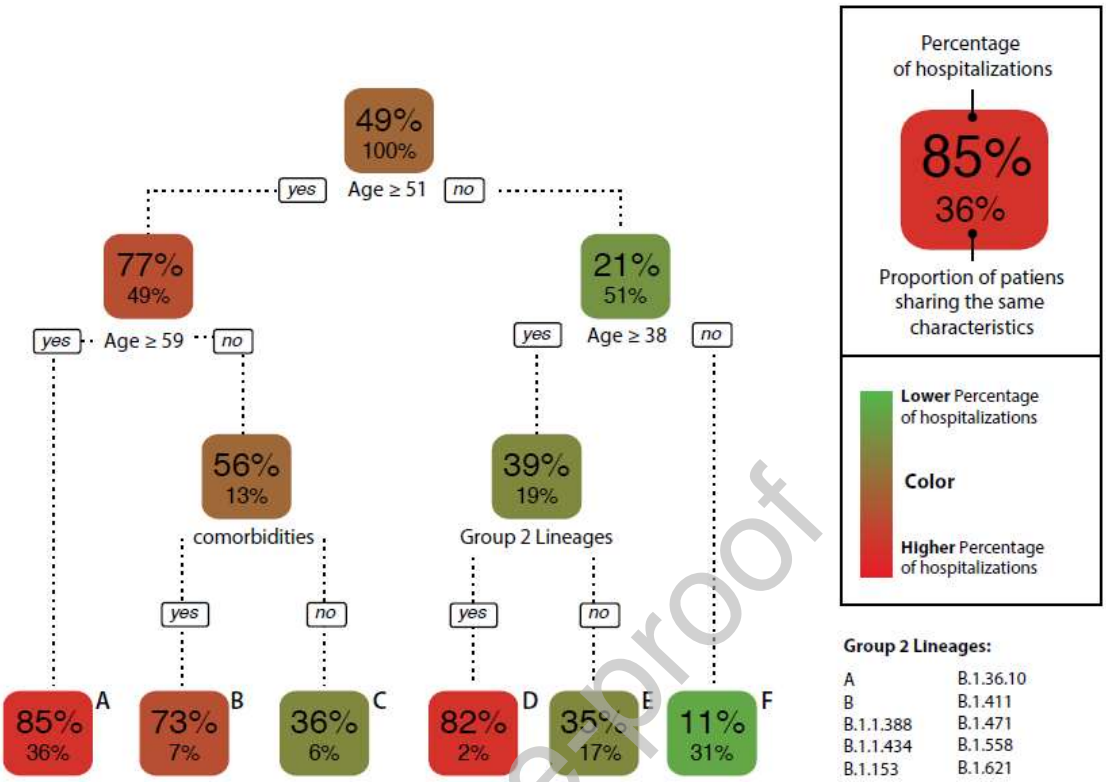

Figure 4. Classification and regression tree for hospitalization

The classification and regression tree for hospitalization seeks to identify the determining variables of COVID-19 clinical severity, in terms of the need for hospitalization. Six groups were identified: A) Patients over 59 years old presenting with 85% hospitalization rate (36% of the sample). B) Patients 51 to 59 years old, with one or more comorbidities, presented 73% hospitalization rate (7% of the sample). C) Patients 51 to 59 years old, but with no comorbidities, presented a substantial reduction in hospitalizations (36%) (6% of the sample). D) Patients 38 to 51 years old with any of the following viral lineages: A, B, B.1.1.388, B.1.1.434, B.1.153, B.1.36.10, B.1.411, B.1.471, B.1.558 or B.1.621, presented 82% hospitalization rate (2% of the sample). E) Patients 38 to 51 years old but not presenting any group D lineage, had 35% hospitalization rate (17% of the sample). G) Patients under 38 years old, had 11% hospitalization rate (31% of the sample).

(This figure should be colored).

Table 1. Baseline demographic characteristics of the cohort

Demographic characteristics of the studied sample n= 997		
	Category	**n(%)**
Age	< 40 years	349 (35%)
	40 - 60 years	312 (31.2%)
	> 60 years	336 (33.7%)
Sex	Female	485 (48.6%)
	Male	512 (51.3%)
Ethnicity	Ethnic majority (Mestizo/White)	764 (76.6%)
	Afro	22 (2.8%)
	Indigenous	5 (0.5%)
	Mulato	12 (1.4%)
	Other	46 (4.6%)
	NA	148 (17.4%)
Nationality	Colombian	987 (98.9%)
	Venezuelan	7 (0.7%)
	Other	3 (0.4%)
Residence	Bogotá	621 (62.2%)
	Cali	289 (29%)
	Cundinamarca	68 (6.9%)
	Other	19 (1.9%)
Age (years)	Mean (SD)	50.6 (18.4)

Table 2. Baseline clinical characteristics of the cohort

Clinical characteristics of the studied sample n= 997		
Variable	**Category**	
Clinical outcomes n (%)	Ambulatory	508 (50.9%)
	In-hospital	328 (32.8%)
	Intensive care	92 (9.2%)
	Deceased	69 (6.9%)
Symptomatology n (%)	Asymptomatic	93 (9.3%)
	Symptomatic	904 (90.7%)
Complications n (%)	Cardiac	48 (4.8%)
	Respiratory	293 (29.4%)
	Renal	79 (7.9%)
	Neurological	36 (3.6%)
	Thromboembolic	28 (2.8%)
Symptoms n (%)	Fever	469 (47.1%)
	Cough	544 (54.5%)
	Fatigue	523 (52.4%)
	Dyspnea	415 (41.6%)
	Diarrhea	173 (17.3%)
	Sore Throat	235 (23.6)
	Anosmia	271 (27.2%)
	Dysgeusia	232 (23.3%)
	Chest pain	110 (11%)
	Nasal congestion	177 (17.7)
	Abdominal pain	62 (6.2%)
	Cyanosis	7 (0.7%)
	Headache	324 (32.4)
BMI	**Mean (SD)**	26.4 (4.4)

Table 3. PANGO lineages detected and their frequencies

PANGO Lineage	WHO label for VOCs / VOIs	n	Percentage (%)
B.1	-	305	45,.0
B.1.111	-	77	11.4
B.1.1.348	-	66	9.7
B.1.153	-	39	5.8
B.1.1	-	39	5.8
B.1.420	-	31	4.6
A	-	16	2.4
B	-	11	1.6
B.1.383[a]	-	9	1.3
B.1.523	-	8	1.2
B.1.621	Mu	7	1.0
B.1.1.388	-	6	0.9
B.1.36.10	-	5	0.7
B.1.293[a]	-	5	0.7
B.1.1.413[a]	-	5	0.7
B.1.1.291[a]	-	5	0.7
B.1.1.28	-	4	0.6
B.1.177.86[a]	-	2	0.3
B.1.1.1	-	2	0.3
B.1.416	-	2	0.3
B.1.1.434[b]	-	2	0.3
B.1.411[a]	-	2	0.3
B.1.36.31[a]	-	2	0.3
B.1.1.213[a]	-	2	0.3
B.1.389[a]	-	1	0.1
B.1.165	-	1	0.1
B.1.1.100	-	1	0.1
B.1.319[b]	-	1	0.1
B.59[b]	-	1	0.1
B.1.456[a]	-	1	0.1
B.1.485[a]	-	1	0.1
B.1.1.409[b]	-	1	0.1
B.1.1.37[a]	-	1	0.1
B.1.505[a]	-	1	0.1

B.1.606[a]	-	1	0.1
B.1.529[a]	-	1	0.1
B.1.565[b]	-	1	0.1
B.1.1.272[a]	-	1	0.1
A.2	-	1	0.1
B.1.281[a]	-	1	0.1
B.1.533[a]	-	1	0.1
B.1.471[a]	-	1	0.1
B.1.324[a]	-	1	0.1
A.2.4[a]	-	1	0.1
B.1.1.7[c]	Alpha	1	0.1
B.1.575[b]	-	1	0.1
B.1.558[a]	-	1	0.1
P.1[c]	Gamma	1	0.1
B.1.1.371[a]	-	1	0.1
B.1.404[a]	-	1	0.1

[a]New in South America; [b]New in Colombia; [c]Variant of Concern

Table 4. Baseline demographic and medical conditions of patients stratified by mortality

Demographics and medical conditions of patients stratified by mortality n= 997				
	Category	**Deceased (n=69)**	**Living (n=926)**	**Chi.Pval**
Age (years) n (%)	< 40 years	3 (4.3%)	346 (37.3%)	3.46E-18
	40 - 60 years	9 (13%)	303 (32.7%)	
	> 60 years	57 (82.6%)	279 (30%)	
Sex n (%)	Female	23 (33 3%)	462 (49.7%)	0.011972952
	Male	46 (66.6%)	466 (50.2%)	
Highest educational level n (%)	Secondary school	44(63.7)	391 (42.1%)	0.000750963
	Professional	25(36.3%)	537 (58%)	
Comorbidities n (%)	Hypertension	38 (55%)	226 (24.3%)	5.38366E-08
	DM2	19 (27.5%)	105 (11.2%)	0.984705638
	Asthma	1 (1.4%)	21 (2.2%)	0.984705638
	COPD	12 (17.3%)	27 (2.9%)	1.47512E-08
	Cardiovascular disease	12 (17.3%)	62 (66.8%)	2.39E-03
	Nephropathy	13 (18.8%)	29 (3.1%)	2.53539E-09
	Malignancy	8 (11.5%)	33 (3.5%)	0.003391579
	Autoimmune disease	3 (4.3%)	24 (2.6%)	0.627410658
	Smoking (formerly)	12 (17.3%)	150 (16.1%)	0.1949283
	Smoking (currently)	0	42 (45.%)	
	Obesity	14 (20.2%)	137 (14.7%)	0.288462289

	HIV	1 (1.4%)	7 (0.7%)	1
	Thyroid disease	11 (15.9%)	82 (8.8%)	0.08123423
Symptoms n (%)	Fever	34 (49.2%)	435 (46.9%)	0.794550071
	Cough	47 (68.1%)	497 (53.5%)	0.026544714
	Fatigue	38 (55%)	485 (52.2%)	0.744480128
	Dyspnea	45 (65.2%)	370 (39.8%)	6.49034E-05
	Diarrhea	10 (14.4%)	163 (17.6%)	0.622079854
	Sore Throat	10 (14.4%)	225 (24.2%)	0.090170707
	Anosmia	5 (7.2%)	266 (28.6%)	0.000201008
	Dysgeusia	4 (5.7%)	228 (24.6%)	0.000643501
	Chest pain	2 (2.8%)	108 (11.6%)	0.041717244
	Nasal congestion	8 (11.5%)	92 (9.9%)	0.809861204
	Abdominal pain	7 (10.1%)	55 (5.9%)	0.253675689
	Cyanosis	1 (1.4%)	6 (0.6%)	0.981464039
	Headache	8 (11.5%)	316 (34%)	0.000207739
	Conjunctivitis	0	25 (2.6%)	0.326209045
	Chills	13 (18.8%)	241 (26%)	0.242789718
Complications n (%)	Cardiac	19 (27.5%)	29 (3.1%)	9.45464E-19
	Respiratory	59 (85.5%)	234 (25.2%)	1.30E-25
	Renal	30 (43.4%)	49 (5.2%)	1.32638E-28
	Neurological	15 (21.7%)	21 (2.2%)	1.04509E-15
	Thromboembolic	7 (10.1%)	21 (2.2%)	0.000575997
BMI (mean, range)		26.127 (16.2-35)	26.437 (16.3-52)	0.588035743

Table 5. Baseline demographics and medical conditions of hospitalized and non-hospitalized patients

Demographics and medical conditions of hospitalized and non-hospitalized patients n= 997				
	Category	**Non-hospitalized (n=508)**	**Hospitalized (n=489)**	**Chi.Pval**
Age (years) n (%)	< 40 years	300 (59%)	49 (10%)	2.90E-75
	40 - 60 years	157 (30.9%)	155 (31.7%)	
	> 60 years	51 (10.1%)	285 (58.3%)	
Sex n (%)	Female	283 (55.8%)	202 (41.3%)	7.31642E-06
	Male	225 (44.2%)	287 (58.7%)	
Highest educational level n (%)	Secondary school	175 (34.4%)	260 (53.1%)	3.75178E-09
	Professional	333 (65.5%)	229 (46.9%)	
Comorbidities n (%)	Hypertension	55 (10.8%)	209 (42.7%)	7.82792E-30
	DM2	27 (5.3%)	97 (19.8%)	7.39187E-12
	Asthma	12 (2.3%)	10 (2%)	0.900344245
	COPD	4 (0.7%)	35 (7.1%)	5.08776E-07
	Cardiovascular disease	10 (1.9%)	64 (13%)	4.87E-11
	Nephropathy	2 (0.3%)	40 (8.1%)	2.51175E-09
	Malignancy	9 (1.7%)	32 (6.5%)	0.000279061
	Autoimmune disease	2 (0.3%)	25 (5.1%)	1.11482E-05
	Smoking (formerly)	86 (16.9%)	76 (15.5%)	6.17715E-05
	Smoking (currently)	35 (6.8%)	7 (1.4%)	
	Obesity	42 (8.3%)	109 (22.2%)	1.15776E-09
	HIV	2 (0.3%)	6 (1.2%)	0.263027525

	Thyroid disease	30 (5.9%)	63(12.8%)	0.000234681
Symptoms n (%)	Fever	196 (38.5%)	273 (55.8%)	7.02857E-08
	Cough	206 (40.5%)	338 (69.1%)	2.39592E-19
	Fatigue	236 (46.4%)	287 (58.6%)	0.000142592
	Dyspnea	98 (19.2%)	317 (64.8%)	9.46952E-48
	Diarrhea	87 (17.1%)	86 (17.5%)	0.913616235
	Sore Throat	151 (29.8%)	84 (17.1%)	4.40343E-06
	Anosmia	222 (43.7%)	49(10%)	1.53071E-32
	Disgeusia	188 (37%)	44 (8.9%)	2.80033E-25
	Chest pain	59 (11.7%)	51(10.4%)	0.62004627
	Nasal congestion	142 (28%)	35(7.1%)	1.77962E-17
	Abdominal pain	29 (5.7%)	33 (6.7%)	0.583361711
	Cyanosis	1 (0.1%)	6 (1.2%)	0.116863239
	Headache	220 (43.5%)	103 (21%)	6.6173E-14
	Conjunctivitis	25 (4.9%)	0	1.882E-06
	Chills	157 (30.9%)	97 (19.8%)	8.24311E-05
Complications n (%)	Cardiac	1 (0.19%)	47 (9.6%)	1.14455E-11
	Respiratory	6 (1.1%)	287 (58.6%)	1.29436E-87
	Renal	0	79 (16.1%)	1.2216E-20
	Neurological	2 (0.3%)	34 (6.9%)	7.43065E-08
	Thromboembolic	0	28 (5.7%)	7.43065E-08
BMI (mean, range)		25.6 (16.3-43)	27.1 (16.2-52.7)	1.88697E-07
Age (years) (mean, range)		39.8 (18.6-92.5)	61.8 (20.7-99.3)	3.14239E-97

Table 6. Demographic Factors and Baseline Clinical Characteristics of the Study Population Stratified by Viral Clade (n=654)

Demographic Factors and Baseline Clinical Characteristics of the Study Population Stratified by Viral Clade (n = 654)						
Age n(%)	**Total (n=652)**	**G (n=167)**	**GH (n=321)**	**GR (n=120)**	**Other (n=46)**	**Chi.Pval**
< 40 years	264 (40.3%)	62 (37.1%)	135 (42%)	60 (50%)	7 (15.2%)	**2.38E-09**
40 - 60 years	206 (31.5%)	48 (28.7 %)	118 (36.7%)	32 (26.6%)	8 (17.3%)	
> 60 years	184 (28.2%)	57 (34.1 %)	68 (21.2%)	28 (23.3%)	31 (67.3%)	
Sex n (%)						
Female	321 (49.1%)	74 (44.3%)	165 (51.4%)	60 (50%)	22 (47.8%)	0.516063827
Male	333 (50.9%)	93 (55.6%)	156 (48.5%)	60 (50%)	24 (52.1%)	
Residence n (%)						
Bogotá	413 (63.2%)	130 (77.8%)	180 (56%)	78 (65%)	25 (54.3%)	1.36E-06
Cundinamarca	40 (6.1%)	15 (8.9%)	13 (4%)	6 (5%)	6 (13%)	
Cali	187 (28.6%)	21 (12.5%)	120 (37.3%)	33 (27.5%)	13 (28.2%)	
Other	14 (2.1%)	1 (0.5%)	8 (2.5%)	3 (2.5%)	2 (4.3%)	
Comorbidities n (%)						
Hypertension	137 (20.9%)	40 (23.9%)	61 (19%)	21 (17.6%)	15 (32.6%)	0.098673612
DM2	71 (10.8%)	19 (11.3%)	28 (8.7%)	13 (10.9%)	11 (23.9%)	0.021689056
Asthma	18 (2.7%)	5 (2.9%)	9 (2.8%)	3 (2.5%)	1 (2.1%)	0.988589637

Cardiovascular disease	45 (6.8%)	15 (8.9%)	18 (5.6%)	8 (6.6%)	4 (8.6%)	0.53035 1676
Nephropathy	24 (3.6%)	8 (4.7%)	6 (1.8%)	5 (4.1%)	5 (10%)	0.01569 4572
COPD	24 (3.6%)	11 (6.5%)	9 (2.8%)	4 (3.3%)	0	0.08998 8724
Malignancy	31 (4.7%)	10 (5.9%)	11 (3.4%)	3 (2.5%)	7 (15.2 %)	0.00250 1822
Autoimmune disease	14 (2.1%)	4 (2.3%)	9 (2.8%)	0	1 (2.1%)	0.34066 3261
HIV	5 (0.7%)	1 (0.6%)	3 (0.9%)	1 (0.8%)	0	0.90896 9613
Smoking (formerly)	118 (18%)	33 (20%)	55 (17.1%)	20 (16.6%)	10 (21.7%)	0.26392 099
Smoking (currently)	25 (3.8%)	8 (4.8%)	11 (3.4%)	3 (9.1%)	3 (6.5%)	0.26392 099
Obesity	93 (14.2%)	24 (14.3%)	41 (12.7%)	14 (11.6%)	14 (30.4%)	0.01114 1853
Symptomatology n(%)						
Symptomatic n (%)	592 (90.5%)	159 (95.2%)	283 (88.1%)	105 (87%)	45 (97.8%)	0.01478 5209
Asymptomatic n (%)	62 (9.5%)	8 (4.7%)	38 (11.8%)	15 (12.5%)	1 (0.2%)	
Symptoms n (%)						
Fever	289 (44.1%)	80 (47.9%)	144 (44.8%)	46 (38.3%)	20 (43.4%)	0.44657 1927
Cough	329 (50.3%)	98 (58.6%)	152 (47.3%)	54 (45%)	26 (56.5%)	0.04889 7461
Fatigue	345 (52.7%)	98 (59.2%)	159 (49.5%)	64 (53.3%)	24 (52.1%)	0.23975 4642
Dyspnea	224 (34.2%)	71 (42.5%)	91 (28.3%)	37 (30.8%)	25 (54.3%)	0.00028 6422
Diarrhea	101 (15.4%)	26 (15.5%)	46 (14.3%)	26 (21.6%)	3 (6.5%)	0.08321 2644
Sore throat	163 (24.9%)	41 (24.5%)	78 (24.2%)	32 (26.6%)	12 (26%)	0.95862 538

Anosmia	207 (31.6%)	47 (28.1%)	107 (33.3%)	46 (38.3%)	7 (15.2%)	0.022403231
Disgeusia	185 (28.2%)	46 (27.5%)	94 (29.2%)	37 (30.8%)	8 (17.3%)	0.350710859
Chest pain	77 (11.7%)	16 (9.5%)	39 (12.1%)	19 (15.8%)	3 (6.5%)	0.267781866
Nasal congestion	132 (20.1%)	38 (22.7%)	60 (18.6%)	32 (26.6%)	2 (4.3%)	0.009658253
Abdominal pain	32 (4.8%)	6 (3.5%)	16 (4.9%)	6 (5%)	4 (8.6%)	0.563171113
Cyanosis	3 (0.4%)	0	1 (0.3%)	0	2 (4.3%)	0.000809846
Headache	223 (34%)	58 (34.7%)	113 (35.2%)	47 (39.1%)	11 (23.9%)	0.332417704
Complications n (%)						
Cardiac	25 (3.8%)	10 (6%)	9 (2.8%)	4 (3.3%)	2 (4.3%)	0.361438087
Respiratory	141 (21.5%)	51 (30%)	50 (15.5%)	22 (18.3%)	18 (39.1%)	2.3284E-05
Renal	41 (6.2%)	17 (10.2%)	11 (3.4%)	7 (5.8%)	6 (13%)	0.005898232
Neurological	16 (2.4%)	4 (2.4%)	6 (1.8%)	3 (2.5%)	3 (6.5%)	0.302293399
Thromboembolic	16 (2.4%)	5 (3%)	4 (1.2%)	4 (3.3%)	3 (6.5%)	0.124023067
Clinical outcomes n (%)						
Ambulatory	409 (62.7%)	83 (49.7%)	231 (71.9%)	88 (73.3%)	7 (15.2%)	1.99E-13
Hospitalized	159 (24.3%)	59 (35.3%)	56 (17.4%)	19 (15.8%)	25 (54.3%)	
ICU	45 (6.8%)	10(6%)	19 (5.9%)	7 (5.8%)	9 (1.9%)	
Deceased	41(6.2%)	15 (9%)	15 (4.6%)	6 (5%)	5 (10.8%)	